electronics

Article

ADQE: Obtain Better Deep Learning Models by Evaluating the Augmented Data Quality Using Information Entropy

Xiaohui Cui [1,2], Yu Li [1,2], Zheng Xie [1,2], Hanzhang Liu [1], Shijie Yang [1] and Chao Mou [1,2,*]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; cuixiaohui@bjfu.edu.cn (X.C.)
[2] Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China
* Correspondence: chao_m@bjfu.edu.cn

Abstract: Data augmentation, as a common technique in deep learning training, is primarily used to mitigate overfitting problems, especially with small-scale datasets. However, it is difficult for us to evaluate whether the augmented dataset truly benefits the performance of the model. If the training model is relied upon in each case to validate the quality of the data augmentation and the dataset, it will take a lot of time and resources. This article proposes a simple and practical approach to evaluate the quality of data augmentation for image classification tasks, enriching the theoretical research on data augmentation quality evaluation. Based on the information entropy, multiple dimensional metrics for data quality augmentation are established, including diversity, class balance, and task relevance. Additionally, a comprehensive data augmentation quality fusion metric is proposed. Experimental results on the CIFAR-10 and CUB-200 datasets show that our method maintains optimal performance in a variety of scenarios. The cosine similarity between the score of our method and the precision of model is up to 99.9%. A rigorous evaluation of data augmentation quality is necessary to guide the improvement of DL model performance. The quality standards and evaluation defined in this article can be utilized by researchers to train high-performance DL models in situations where data are limited.

Keywords: data augmentation; deep learning; data quality; big data; data mining

Citation: Cui, X.; Li, Y.; Xie, Z.; Liu, H.; Yang, S.; Mou, C. ADQE: Obtain Better Deep Learning Models by Evaluating the Augmented Data Quality Using Information Entropy. *Electronics* **2023**, *12*, 4077. https://doi.org/10.3390/electronics12194077

Academic Editor: Byung-Gyu Kim

Received: 21 August 2023
Revised: 16 September 2023
Accepted: 27 September 2023
Published: 28 September 2023

1. Introduction

Data-driven deep learning (DL) has achieved many significant achievements in the past few years [1–3], and data augmentation has played an important role in it [4–6]. In practical applications, often the scale of the available data is insufficient for model training, which is the problem that data augmentation aims to solve [7]. Data augmentation is a regularization technique that increases the size and diversity of datasets by transforming and augmenting the original data [8]. Typically, data augmentation has a positive effect on the training and performance of DL models, but in practice, phenomena such as decreased precision or overfitting can occur after data augmentation [9]. Therefore, we need an evaluation criterion for the quality of data augmentation to assess the effectiveness of the employed data augmentation methods and their ability to improve model performance and generalization. Most of the research on data augmentation focuses on enhancing the generalization characteristics of the models, such as precision and F1 score [10,11]. Only a few papers have proposed a general framework for explaining data augmentation by studying its regularization effects [12–14], its influence on feature selection [15–17], rough set [18–20], and invariance perspectives [21,22]. However, the evaluation methods based on model performance cannot explain in which aspect data augmentation improves data quality. Neither can it guide researchers to choose a more optimal data enhancement strategy, even if a lot of time is spent on training the model. In order to determine to what extent data augmentation can improve data quality, evaluation standards need to be more

theoretical and comprehensive. By visualizing and analyzing the augmented data, we can evaluate the effectiveness of data augmentation, observe whether the changes in data are reasonable, and cover different categories and tasks. This intuitive and simple method can efficiently evaluate the effects of data augmentation and quickly find suitable data augmentation methods, which can help obtain the most suitable dataset for model training in advance.

Studies [23,24] have shown that data quality is a multidimensional concept. Data quality has different meanings in different contexts. For example, data quality can be about measuring defective or outlier data in a general context [25–27], or describing whether the data meet the expected purpose in a specific context [28]. In this paper, we define data quality as a measure of data suitability for constructing a DL training set. Existing data quality assessments consider both intrinsic data quality and contextual quality [29], but the definitions of contextual quality vary. The most common idea is to divide contextual quality into two parts based on the process of DL: diversity within the training set and similarity between the training and testing sets. The main idea is to make the training set complex enough to encompass all the features and be similar to the real distribution represented by the testing set so that the DL model can learn adequately from this dataset. However, they overlook the fact that the performance of deep learning models is not only influenced by the problem space covered by the data. For instance, imbalanced classes in the dataset may lead to model bias [30,31], and these imbalances can occur in terms of quantity, features, or colors. Creating more dimensions based on the task and data features can better describe the quality of the data and its value for deep learning models. We hope to construct a universal, robust, and highly generalizable multidimensional quality evaluation method by refining and differentiating the definition of quality metrics, which can provide strong support for the quality evaluation of data augmentation.

In addition, due to the curse of dimensionality, such as in the case of image and text data, there arise computational and statistical challenges, with computational complexity growing exponentially. Hence, many works have used average similarity and minmax similarity between samples to calculate these two dimensions [29,32]. Although average or minmax similarity between samples can quickly assess the quality of a dataset, they cannot accurately approximate the precision of models trained on that dataset. Information entropy [33] can provide a comprehensive evaluation of data distribution, considering global characteristics such as sample diversity, rather than just focusing on average differences in the data [34]. Its feature as a non-linear measure based on probability distribution can better capture non-linear relationships in data distribution, with less sensitivity to noise and stronger interpretability [35]. Because of the low noise sensitivity, it is suitable to improve the computational efficiency with dimension reduction technology. The computation problem caused by a dimension disaster can be avoided. In summary, information entropy, as a metric for evaluating the quality of data augmentation, possesses more comprehensive, robust, and interpretable characteristics, making it more suitable for approximating the precision of models.

Therefore, this paper proposes an information entropy-based method for evaluating the quality of data augmentation. By attempting to deconstruct the dimensions of the data, we assess the quality of the dataset and data augmentation. In our approach, the augmented dataset is initially broken down into three dimensions, including diversity, class balance, and task relevance. Furthermore, taking image data as an example, for each dimension, numerous sub-dimensions are derived based on the task and data characteristics. Finally, by considering the correlations between the metrics, we calculate the ultimate composite metric score, providing insights into the impact of the current augmentation strategy on model performance.

- In this paper, we design and implement a data enhancement quality evaluation method, which can optimize and generate large-scale, high-quality data sets by disassembling and balancing the quality dimensions of data sets.

- This paper discusses the choice of mathematical tools for statistical analysis of data dimensions, and determines that information entropy is more suitable than other methods for evaluating the information content of data.
- This paper extensively evaluates the proposed method on various data augmentation techniques, datasets, and models. There is a strong correlation between the experimental results of the deep learning model and the evaluation results of the method, which shows that the method can improve the performance of the model on related tasks by evaluating the data enhancement quality.

2. Methods

In this work, we aim to explore the effectiveness of data augmentation in enhancing datasets, with the hope of replacing expensive model training with more comprehensive statistical metrics to evaluate the quality of augmented datasets. The primary goal of data augmentation is to generate a diverse and balanced dataset that is highly relevant to the task.

2.1. Preliminaries

Before presenting the details of our method, we give a brief overview of deep learning and data augmentation, which provides the theoretical basis of our algorithm design. For better illustration, some notations are summarized in Table 1.

Table 1. Some important mathematical notation.

Mathematical Notation	Description
X, Y	The data and labels of the dataset, as well as the input and output space of the model. X and Y represent the original training dataset. X' and Y' represent the augmented training dataset. X_t and Y_t represent the test dataset.
x, y	The x denotes an input data and the y denotes a label for the data. The x and y represent the original training dataset. The x' and y' represent the augmented training dataset. The x_t and y_t represent the test dataset.
P, p	Both represent a probability function that describes the distribution of the sample space.
R	It represents the risk function. Its subscripts represent the computational ideas used, empirical and expectation, respectively.
Q	It represents a collection of data augmentation quality metrics. Each $Q_i \in Q$ counts the different dimensions of the data.
pixel	It represents the pixel of the image data.
D	D denotes the dataset. D represents the original training dataset. D' represents the augmented training dataset. D_t represents the test dataset.
C, c_i	C represents the number of classes in the dataset and c_i represents the number of samples in the ith classes of the training dataset.
N	It represents the number of samples in the dataset. N represents the original training dataset. N' represents the augmented training dataset. N_t represents the test dataset.

2.1.1. Deep Learning

In machine learning, we formally refer to the sets of all possible values for the input and output of models as the input space X and the output space Y, respectively. Each specific input is an instance x, usually represented by a feature vector. In this case, the space where all feature vectors exist is called the feature space. The specific output is denoted as y, typically representing the label of the input X. At this point, the input and output variables X and Y follow a joint probability distribution $P(X, Y)$. The input space X and the output space Y together form a sample space. For a sample $(x, y) \in (X, Y)$ in the sample space, it is assumed that there exists an unknown true mapping function $f : X \to Y$, such that $y = f(x, \theta)$, where $\theta \in R^m$ represents the parameters in the function space and m is the number of parameters. In this case, we can measure the distance between the model

and the data distribution $P(X, Y)$ to train the model. Therefore, the expected loss of the model $f(x, \theta)$ with respect to the joint distribution $P(X, Y)$ is expressed as

$$R_{exp}(\theta) = E_P[L(y, f(x, \theta))] = \int_{X \times Y} L(y, f(x, \theta)) P(x, y)\, dx\, dy, \tag{1}$$

since the $L(y, f(x, \theta))$ represents the loss function, which quantifies the difference between individual input and output instances. However, in reality, the data distribution $P(X, Y)$ is often unknown, and we only have knowledge of the distribution of samples in the training set. Therefore, to deal with this situation, in DL, the approach is to minimize the expected loss on the training set. As shown in Equation (2), the empirical distribution $\hat{P}(X, Y)$ based on the training set is used instead of the true distribution $P(X, Y)$ to calculate the empirical loss R_{emp}. This way, during the training process, the model performs parameter optimization based on the sample distribution in the training set, aiming to approximate the performance of the true distribution as closely as possible.

$$R_{emp}(\theta) = \frac{1}{N} \sum_{n=1}^{N} L(y, f(x, \theta)). \tag{2}$$

2.1.2. Data Augmentation

Data augmentation refers to any method that uses artificial transformations of data and labels to expand the original training set. It can be represented as a mapping of the set. This function is defined as

$$P(X^{'}, Y^{'}) = f((X, Y)|\beta) \bigcup P(X, Y), \tag{3}$$

where β represents the data augmentation strategy, $f((X, Y)|\beta)$ represents the augmented part, and $P(X^{'}, Y^{'})$ represents the final augmented training dataset.

2.1.3. Benefits of Data Augmentation

To address the issue of unreliable empirical risk in situations with limited data, we need to introduce prior knowledge. Prior knowledge is used to augment the dataset, even in cases where data are scarce.

Lemma 1 (Chebyshev's inequality)**.** *Let t be a random variable with finite expected value μ. And there are n variables in total. Then, for any small positive number ϵ,*

$$\lim_{n \to \infty} P(|\frac{\sum t_i}{n} - \mu < \epsilon|) = 1. \tag{4}$$

According to the law of large numbers, Equations (2) and (4), as the sample size N becomes sufficiently large, the empirical risk R_{emp} tends to the expected risk R_{exp}. However, when it comes to DL datasets, considering only the distribution of samples and labels is insufficient. For instance, simply resampling data can lead to more severe overfitting of the model. We also need to ensure that the features in the data align closely with the true distribution, which is a key problem addressed by data augmentation. In general, data augmentation is achieved by modifying the original data based on prior knowledge to expand the dataset. The generated data may have the same labels as the original data, but the features extracted by the model are different. This enables the model to more easily recognize the critical features relevant to the task at hand. The parameters for data enhancement need to satisfy the following expression:

$$\arg\min_{\beta} distance(\arg\max_{P()} quality(P(X^{'}, Y^{'})), P(X_t, Y_t)), \tag{5}$$

where the *distance* is defined as a function of measuring distribution distance, such as similarity, and *quality* is defined as a function that measures the distribution of data sets.

The ideal scenario is that the dataset remains consistent with the true distribution for all features $P(X_t, Y_t)$. However, this is an ideal situation and the true distribution is still unknown. Therefore, we use the test set instead of the true distribution for the estimation.

Theorem 1. *The expectation and variance of the original dataset are μ and σ, the expectation and variance of the augmented dataset are $\mu^{'}$ and $\sigma^{'}$. Assuming that the expectation and variance of true distribution are μ_t and σ_t. Equation (5) can be expressed as*

$$fusion(distance(\mu^{'}, \mu_t), distance(\sigma^{'}, \sigma)). \tag{6}$$

However, due to the randomness of data augmentation, the generated data may not necessarily be more in line with the true distribution compared to the original data. Therefore, we need to perform quality estimation on it.

Lemma 2. *Expectation and variance are not mutually independent, only when the distribution does not follow a normal distribution.*

Therefore, the quality evaluation metric for data augmentation is defined as

$$Q_{Augmentation} = Q_\sigma \times Q_\mu = Q_\sigma \times Q_{TaskRelevance}, \tag{7}$$

which is the product of the statistical values of expectation Q_σ and variance Q_μ . The expectation of the dataset is primarily influenced by the target task, and the value obtained from expectation is also referred to as task relevance.

Unlike expectation values, data can have different distributions based on different feature selections, leading to varying variances. These variances can be mainly divided into two categories. One is the distribution of semantic features that approximate a normal distribution, and the other is the distribution of categories, which is mostly a uniform distribution. The variances of these two distributions are independent of each other, so statistical values of variances can be obtained using addition.

Theorem 2. *Output space $Y = (y_1, y_2, \cdots, y_n) \sim U(\lambda)$, input space X is abstracted into $feature = (feature_1, feature_2, \cdots .feature_{k_1}) \sim N(\mu_1, \sigma_1)$. So $P(x, y)$ can be represented by $P(feature, label)$. By delineating the distribution of label-independent features, we have the $semantic_i = (semantic_1, semantic_2, \cdots, semantic_{k_2}) \sim N(\mu_2, \sigma_2), semantic_{k_2} \in C$, then*

$$P(label, semantic) = \sum_{i}^{C} P(label) \times P(semantic_i) = \sum_{i}^{C} P(semantic_i). \tag{8}$$

Features are classified according to whether they have a relationship with the category, and the same is true for data quality. So we have

$$Q_\sigma = (Q_{label} + Q_{(label,feature)}) + Q_{semantic} = Q_{ClassBalance} + Q_{Diversity}. \tag{9}$$

Finally, the augmented datasets quality evaluation formula is defined as

$$Q_{Augmentation} = (Q_{Diversity} + Q_{ClassBalance}) \times Q_{TaskRelevance}. \tag{10}$$

2.2. The Overall Framework

The overall framework of the method proposed in this paper is shown in Figure 1 and Algorithm 1. The methodology of this paper will consist of the following components: (1) Feature Extraction: The dataset will be processed using different data augmentation strategies. And each sample in the datasets will be mapped to a high-dimensional feature space. (2) Metric Scores Calculating: The quality metrics of the dataset will be divided into three dimensions: diversity, class balance, and task relevance. Information entropy is

primarily used to evaluate the diversity and class balance metric. (3) Result Statistics: The scores of each augmented dataset will be ranked, and the datasets with higher quality will be selected for model training and validation. The implementation code is available at: https://github.com/ForestryIIP/ADQE (accessed on 20 August 2023).

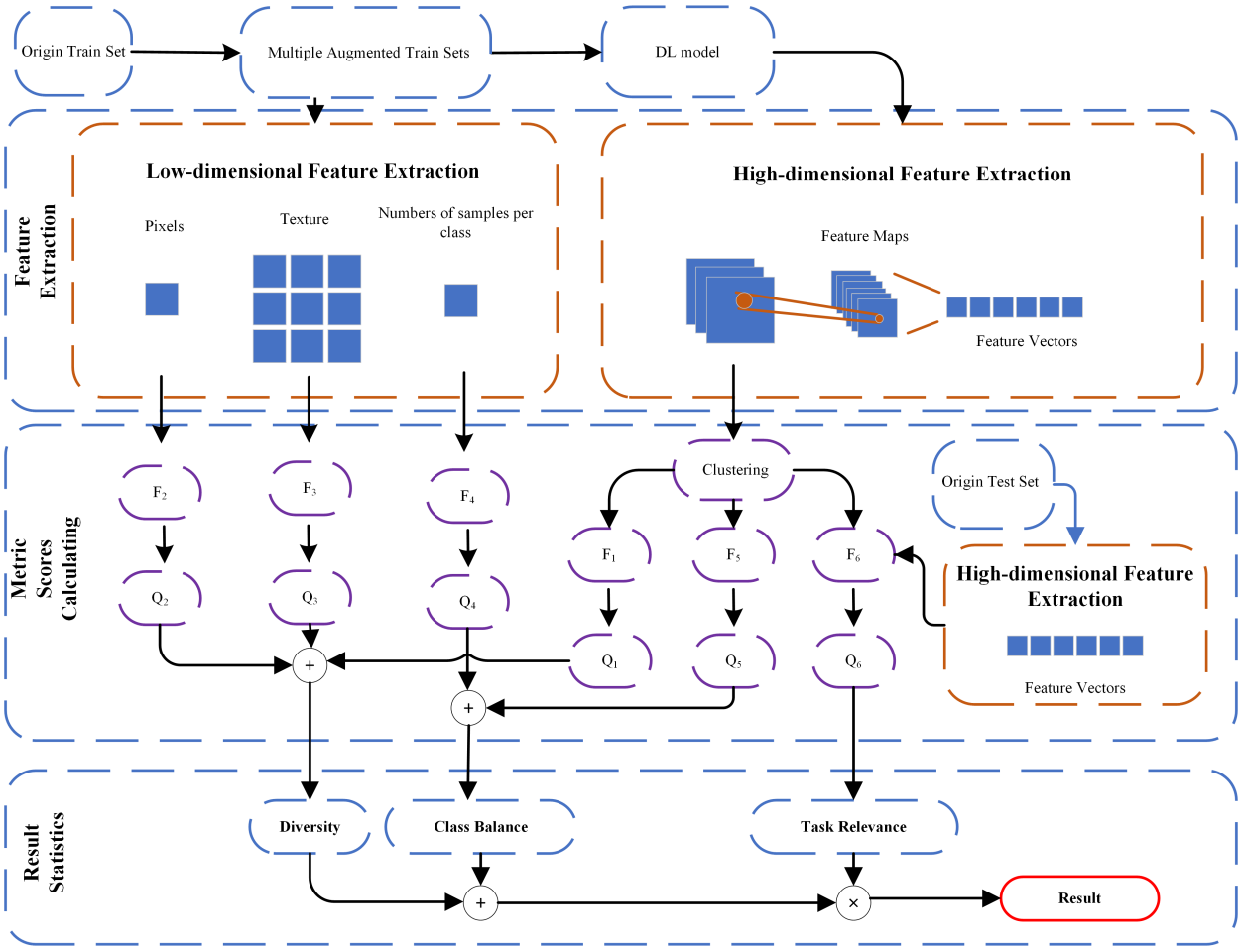


Figure 1. The framework of the proposed method.

Algorithm 1: Calculation of all augmented datasets evaluations.

Data: The collection of augmented datasets $D^{'} = \{D_0^{'}, D_1^{'}, \cdots, D_s^{'}\}$, A test dataset D_t, DL model NN

Result: the quality metric of the augmented dataset Q

$Q \leftarrow \emptyset$;
$cnt \leftarrow 0$;
for *each dataset* $D_i^{'} \in D^{'}$ **do**
  Update Q_4 *based on Algorithm* 4;
  Update Q_2 *and* Q_3 *based on Algorithm* 3;
  Extracting the feature vector V from $D_i^{'}$ with help of NN;
  Clustering feature vector V into k classes of $V^{'}$;
  Update Q_5 *based on Algorithm* 5;
  Update Q_1 *based on Algorithm* 2;
  Update Q_6 *based on Algorithm* 6;
  if $cnt\,!=\,0$ **then**
    $Q \leftarrow$ *result of Equation* (21)
  end
end

2.3. Feature Extraction

2.3.1. Feature Selection

The selection of features is crucial for evaluating the quality of data augmentation, as it impacts the effectiveness and comprehensiveness of metric calculations. In order to obtain more effective image representations, this paper utilizes the pre-trained model ResNet101 [36] to map images into high-dimensional feature vectors. Additionally, to achieve more comprehensive metrics, we supplement the quality indicators by incorporating features such as pixel brightness, texture, and class sample counts based on the characteristics of the images and the dataset itself.

2.3.2. Clustering

After data augmentation, we can proceed with the calculation of various metrics. For diversity computation, calculating low-dimensional feature diversity simply requires measuring the individual feature values of each image, which can be performed in O(n) time complexity. However, in the case of high-dimensional features, the complexity increases as we need to calculate the similarity between each pair of samples and compute the eigenvalues of the nn similarity matrix, resulting in a time complexity of O(n^3). Similarly, for task relevance, we need to compute the similarity between sample pairs from the training set and the test set, resulting in a high time complexity of O($n \times m$). To address these challenges, this study adopts the FINCH clustering algorithm [37], which employs pooling and sampling techniques to reduce the dimensionality and scale of the dataset, thereby mitigating the computational cost and time overhead. In the calculation of class balance, the clustering algorithm can directly compute the desired metrics.

2.4. Metric Scores Calculating

2.4.1. Diversity

The main benefit of data augmentation comes from increasing diversity. However, only comparing the average similarity of data in the augmented dataset in high-dimensional space has limitations. It just only indicates the pairwise distances between data points and does not reflect the overall distribution of data in the high-dimensional feature space. The Vendi score [38] introduces an ecological concept by using the entropy index of species distribution. This effectively measures the diversity of feature distributions in the training set samples. The Vendi score is defined as the exponential Shannon entropy of the eigenvalues of a similarity matrix K, such as

$$K(x,y) = \{similarity(x,y)|x,y \in D^{'}\}. \tag{11}$$

This matrix is derived from a user-defined similarity function applied to the samples under evaluation for diversity:

$$Q_1 = \frac{1}{C}\sum_{i=1}^{C} exp(-\sum_{j=1}^{c_i} \lambda_j log\lambda_j), \tag{12}$$

where the λ_j represents the eigenvalues of the similarity matrix K. The K is a positive-definite matrix obtained from a set of samples x and the similarity function. For all x, $similarity(x,x) = 1$. This method , as shown in Algorithm 2, primarily quantifies the effective number of distinct elements in the data. For example, after extracting features from an image using a neural network, you typically obtain a 2048-dimensional vector. This vector stores features across different dimensions of the image. Assuming the similarity function is a cosine similarity, this metric measures whether the directions of two vectors align. If the feature dimensions forming these directions are more similar, the metric value is higher. Consequently, a similarity matrix K can be computed. Eigenvalues generally represent inherent structural properties and patterns within the data [39]. Each eigenvalue corresponds to a mode of variation or structure in the data. In the context of image similarity,

they can indicate different similarity patterns or clusters among images. The magnitude of eigenvalues also reflects the proportion of corresponding patterns in the data. Therefore, computing the entropy of eigenvalues is equivalent to quantifying the richness of patterns in the data. If a single pattern dominates the data, the style and content of the images are highly certain, resulting in low information uncertainty. Conversely, when multiple similar patterns share similar proportions, the style and content of the images become less certain, leading to higher information uncertainty. In summary, the eigenvalues of similarity matrix K and the entropy of these eigenvalues provide valuable insights into the data's structure and diversity. High entropy indicates complexity and diversity, whereas low entropy suggests simplicity or uniformity in similarity patterns within the data. They can guide overall data analysis in the context of diversity metrics.

Algorithm 2: Rapid calculation of diversity by clustering.

Data: A training dataset $D^{'} = \{X_1, X_2, \cdots, X_{N'}\}$, dataset size $N^{'}$, DL model NN
Result: Diversity Metric Scores Q_1

$K \leftarrow 0_{[N' \times N']}$;
$Q_1 \leftarrow 0$;
Extracting the feature vector V from $D^{'}$ with help of NN;
Clustering feature vector V into k classes of $V^{'}$;
$D^{'}_{clustered} \leftarrow \emptyset$;
for $i \leftarrow 0$ *to* k **do**
    $D^{'}_{clustered} \leftarrow$ randomly select C data from different categories in $V^{'}$
end
for *class index* $\leftarrow 0$ *to* C **do**
    for *each vector* v_i *and* $v_j \in$ *collection of class index in* $V^{'}$ **do**
        $K[i][j] \leftarrow similarity(v_i, v_j)$;
    end
    $\lambda \leftarrow eigenvalue\,of\,K$;
    $Q_1 \leftarrow Q_1 + exp(-\sum_{j=1}^{S_i} \lambda_j log \lambda_j)$;
end
$Q_1 \leftarrow Q_1 \div C$;

Ignoring the time of feature extraction and clustering part, the computation of Q_1 is mainly divided into two steps. They are the computation of similarity adjacency matrix and its eigenvalues, respectively. For each pair of vectors, similarity computation, such as cosine similarity, requires $O(K)$ time complexity. Where K is the dimension size of the image vector output by the model. Every element in the dataset needs to be computed, then computing the adjacency matrix requires $O(KN^2)$ time complexity. The next step is to solve for the eigenvalues, which are also usually $O(N^3)$. Generally, $K \ll N$, so the time complexity is $O(N^3)$. Although the time complexity is high, the data size is reduced by clustering in advance. The actual computation time is still within acceptable limits.

Furthermore, analyzing the diversity of data solely based on high-dimensional features provides an abstract understanding of diversity, but it may not provide an intuitive and clear understanding. Therefore, it is still necessary to define the diversity of data in low-dimensional features. For example, in image data, models often learn texture features extensively from the dataset [40]. To address this, we can calculate the occurrence probabilities of different textures in all the data:

$$Q_2 = exp(\sum_{i=1}^{N_{Texture}} p_{Texture_{i,j}} log p_{Texture_{i,j}}), \tag{13}$$

where $N_{Texture}$ is the number of different $Texture$ and $p_{Texture_{i,j}}$ represents the probability that the value of $Texture_{i,j}$ will occur. The Texture is defined as a combination of $pixel_{i,j}$ and adjacent pixels:

$$Texture_{i,j} = \begin{pmatrix} pixel_{i-1,j+1} & pixel_{i,j+1} & pixel_{i+1,j+1} \\ pixel_{i-1,j} & pixel_{i,j} & pixel_{i+1,j} \\ pixel_{i-1,j-1} & pixel_{i,j-1} & pixel_{i+1,j-1} \end{pmatrix}. \tag{14}$$

Due to computational complexity and memory limitations, this paper calculates the information entropy of the average pixel values within a 3×3 window instead of the information entropy of pixel combinations. Lastly, considering that brightness can have an impact on the model [41], this paper also calculates the probability of brightness for each pixel in the entire dataset and computes its entropy value:

$$Q_3 = \frac{1}{3} \sum_{RGB=1}^{3} exp(\sum_{level=1}^{256} p_{level} log p_{level}), \tag{15}$$

where RGB represents the three color channels of an image and $level$ represents the component of the channel, up to a maximum of 256. Texture and brightness are counting the pixels of the image, so the time complexity is $O(PN)$ and P is the average number of pixel points in the image. As shown in Algorithm 3, the variables for the Q_2 and Q_3 are similar and can be run together in the feature extraction phase.

Algorithm 3: The calculation of image datasets' color and brightness space.

Data: A training dataset $D^{'}$
Result: Diversity Metric Scores Q_2 and Q_3

$P_1 \leftarrow 0_{[3\times256]}$;
$P_2 \leftarrow 0_{[3\times256]}$;
for *each image* $x \in D^{'}$ **do**
  for $channel_{RGB} \leftarrow 0$ *to* 2 **do**
    for *each pixel* $\in x$ **do**
      $t \leftarrow Average\ of\ adjacent\ pixels$;
      $P_1[channel_{RGB}][pixel] \leftarrow P_1[channel_{RGB}][pixel] + 1$;
      $P_2[channel_{RGB}][t] \leftarrow P_2[channel_{RGB}][t] + 1$;
    end
  end
end
$Q_2 \leftarrow 0$;
$Q_3 \leftarrow 0$;
for $channel_{RGB} \leftarrow 0$ *to* 2 **do**
  $P_1[channel_{RGB}] \leftarrow sum(P_1[channel_{RGB}])$;
  $Q_2 \leftarrow Q_2 + exp(-\sum_{j=0}^{255} P_1[channel_{RGB}][j] log(P_1[channel_{RGB}][j]))$;
  $P_2[channel_{RGB}] \leftarrow sum(P_2[channel_{RGB}])$;
  $Q_3 \leftarrow Q_3 + exp(-\sum_{j=0}^{255} P_2[channel_{RGB}][j] log(P_2[channel_{RGB}][j]))$;
end
Calculate entropy;

2.4.2. Class Balance

From the perspective of data categories, real-world datasets often suffer from the long-tail problems [42]. Models tend to overlook minority classes due to the majority of samples being concentrated in a few categories, resulting in poorer predictive performance. To measure class balance, the first aspect to consider is whether the number of samples

in each class is equal, which is the most fundamental metric. The formula for balance of number in class is as follows:

$$Q_4 = 1 - \frac{1}{C}\sum_{i=1}^{C}(c_i - \bar{c}), \tag{16}$$

where the $\bar{c}$ represents the average count of samples across classes. The variance of the classes is computed with a time complexity of $O(C)$.The algorithm is illustrated in Algorithm 4.

Algorithm 4: The calculation of quantity balance between classes.

Data: A training dataset $D^{'}$, The number of classes C
Result: Class Balance Metric Scores Q_4
$K \leftarrow 0_{[C\times 1]}$;
for $i \leftarrow 0$ *to* C **do**
  $K[i] \leftarrow c_i$
end
$Q_4 \leftarrow 1 - \frac{1}{C}(\sum_{i=1}^{C}(K[i] - average(K)))$;

Additionally, we need to assess the distribution differences between classes. When performing a classification task, it is generally easier to distinguish between animals and humans compared to distinguishing between males and females. If both labels are present in the same dataset, the labels for males and females can easily be confused. This results in poor data quality because the model cannot learn precise features to differentiate between males and females accurately [43]. Unsupervised clustering can group similar samples into the same class. In a dataset with balanced difficulty and granularity of classification, the clustering results should align with the original labels. By calculating the mutual information between the clustering results and the original labels, we can obtain the classification difficulty of the samples, which serves as an indicator for the class feature balance of the dataset. The basic definition of information entropy is

$$H(X) = -\sum_{X} p(x) log p(x), \tag{17}$$

where $H(X)$ represents the entropy of the classes. The basic definition of mutual information is

$$I(X;Y) = \sum_{X}\sum_{Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}, \tag{18}$$

where $I(X;Y)$ represents the mutual information and $p(x,y)$ is the joint distribution between the clustering results and the ground truth. The $p(x)$ and $p(y)$ are the marginal distributions of the ground truth labels and clustering results, respectively. Although mutual information can also measure the degree of similarity between two clustering results, its value is strongly influenced by the sample size. Normalized mutual information (NMI), on the other hand, can better measure the degree of similarity between two clustering results by normalizing the mutual information values to the same range of values. The NMI is defined as

$$Q_5 = 2\frac{I(X;Y)}{H(x) + H(y)}. \tag{19}$$

The details are presented in Algorithm 5. Ignoring the time of feature extraction and clustering part, the computational time complexity of the mutual information of the label distributions before and after clustering is only $O(N)$.

Algorithm 5: The calculation of features balance between classes.

Data: A training dataset $D^{'}$
Result: Class Balance Metric Scores Q_5
Extracting the feature vector V from $D^{'}$ with help of NN;
Clustering feature vector V into k classes of $V^{'}$;
$P_{D'} \leftarrow distribution\ of\ V$;
$P_{D'_{clustered}} \leftarrow distribution\ of\ V^{'}$;
$P_{D',D'_{clustered}} \leftarrow jointdistribution\ between\ V and\ V^{'}$;
Update Q_5 based on Equations (18) and (19);

2.4.3. Task Relevance

The most common issue that can occur after image augmentation is the loss of semantic information, where important features, such as the pixels representing a dog in a dog image, are completely erased. In such cases, even humans would struggle to determine the correct label for the modified image. Therefore, assessing the relevance between the augmented training set and the task at hand can provide insights into the quality of the dataset. Considering that models are typically evaluated on a separate test set for various tasks, this study compares the average similarity between the test set and the augmented training set in the high-dimensional feature space:

$$Q_6 = \frac{1}{N_t * N'} \sum_{i=1}^{N_t} \sum_{j=1}^{N'} similarity(i,j), \tag{20}$$

where $similarity(i,j)$ denotes the similarity function used to measure the similarity between samples. Based on the Equation (20), we can get the Algorithm 6.

Algorithm 6: Rapid calculation of task relevance by clustering.

Data: A training dataset $D^{'} = \{x_1, x_2, \cdots, x_{N'}\}$, dataset size $N^{'}$, The number of classes C, A test dataset D_t, test dataset size N_t, DL model NN
Result: Task Relevance Metric Scores Q_6
$Q_6 \leftarrow 0$;
Extracting the feature vector V from $D^{'}$ and V_t from D_t with help of NN;
Clustering feature vector V into k classes of $V^{'}$;
$D^{'}_{clustered} \leftarrow \emptyset$;
for $i \leftarrow 0\ to\ k$ **do**
    $D^{'}_{clustered} \leftarrow$ randomly select C data from different categories in $V^{'}$
end
for $class\ index \leftarrow 0\ to\ C$ **do**
    for *each vector* v_i *and* $v_j \in$ *collection of class index in* $V^{'}$ *and* V_t **do**
        $Q_6 \leftarrow Q_6 + similarity(v_i, v_j)$;
    end
end
$Q_6 \leftarrow average(Q_6)$;

Ignoring the time of feature extraction and clustering part, the computation of Q_6 is mainly divided into two steps. They are the computation of the similarity matrix and its average, respectively. The similarity of all unordered pairs of the dataset and test set needs to be computed, assuming that the dataset size is N and the test set size is M, then the total number is $N \times M$. The similarity matrix time complexity is $O(KMN)$. Where K is the

dimension size of the image vector output by the model. And calculating the mean only requires $O(NM)$. So the total time complexity is $O(KMN)$.

2.5. Result Statistics

To combine various metrics, existing approaches often employ weighted fusion [44] or rank fusion [45] to directly rank the augmented datasets. However, these methods are not suitable in this case because the metrics from different dimensions cannot be directly compared, and a comparison between the augmented training set and the original training set is required. In this paper, a fusion approach is proposed that calculates the quality of data augmentation by considering the ratio of quality before and after augmentation. All metrics are shown in Table 2 and the fusion equation is defined as

$$Q = \frac{Q_6^{'}}{Q_6} \times \sum_{i=1}^{5} w_i \frac{Q_i^{'}}{Q_i} / \sum_{i=1}^{5} w_i, \tag{21}$$

where the $Q_i^{'}$ represents the quality metric of the augmented dataset, Q_i represents the quality metric of the original dataset and w_i represents the weight of each metric to the final fusion metric Q. By assigning appropriate weight coefficients to the metrics of different parts, the balance of influence of different factors in the metrics can be ensured. For example, when you encounter high precision requirements of the task, task relevance metrics are more critical and need to be assigned a higher weight.

Table 2. Data augmentation quality metrics.

Quality Metrics	Full Name	Category
Q_1	diversity of features	diversity
Q_2	diversity of textures	
Q_3	diversity of brightness	
Q_4	balance of number between class	class balance
Q_5	balance of features between class	
Q_6	task relevance	task relevance

3. Experimental

3.1. Datasets and Data Augmentation

In order to validate the effective evaluation of data augmentation quality, this study selected different augmentation strategies, datasets at different granularities, and several specific tasks as the objects of method evaluation. The CIFAR-10 and CUB-200 [46] datasets were chosen as the experimental datasets for image classification tasks. The CIFAR-10 and CUB-200 represent different areas of computer vision problems. CIFAR-10 is an image classification dataset containing 10 different categories of common objects such as aircraft, dogs, cars, and so on. The CUB-200, which focuses on bird identification, contains images of 200 different species of birds. The multi-domain coverage of these two datasets allowed us to explore the impact of diversity and task relevance in different application contexts. They also exhibit varying levels of image diversity and class balance. CIFAR-10 includes diverse scenes, lighting conditions, angles, and variations. In contrast, CUB-200 exhibits limited image diversity, with predominantly consistent backgrounds. Furthermore, both datasets represent numerous practical application scenarios, such as image classification, object detection, and object recognition. By conducting experiments on CIFAR-10 and CUB-200, we gain a better understanding of the performance and applicability of our methods.

In the context of image data applications, data augmentation methods add noise to the original data to simulate other real-world scenarios, thus creating augmented images for model training. To evaluate the improvement in data quality brought by different data augmentation methods, this study employs RandAugment's data augmentation search strategy [47]. Unlike RandAugment, which applies random transformations to images during training, this study scales up the dataset by employing the RandAugment

strategy prior to training. To generate data diversity, this study selects n transformations from a set of $k = 16$ data augmentation transformations with uniform probability, where the augmentation magnitude for each transformation is set to m. By varying these two parameters, the strategy can express a total of $m \times 2^n$ potential augmentation policies, where n represents the strategy for selecting data enhancement and m represents the intensity of data augmentation. Each parameter of the transformations is scaled using the same linear scale, ranging from 0 to 30, where 30 represents the maximum scale for a given transformation, and then mapped to the parameter range of each transformation. Subsequently, during the expansion and generation of the augmented dataset, we uniformly sample dataset samples with probability c for transformation. All generated images are then merged with the original dataset to create the augmented dataset. However, image data alone is insufficient for calculating quality metrics. Therefore, the image data needs to be abstracted into 2048-dimensional feature vectors. In this study, a pre-trained model, ResNet101, is utilized to extract features from the generated augmented dataset.

3.2. Baseline

In this paper, we will present the results of our proposed method on CIFAR-10 and CUB-200 datasets to demonstrate how our approach captures the intuitive notion of data augmentation and can be applied to assess the quality of data augmentation in DL. The existing data augmentation evaluation work can be divided into two categories: model validation for assessing effectiveness and statistical analysis for improving data quality. Model validation aims for the highest level of accuracy, as seen in approaches like AutoAugment. However, due to the immense computational demands, it may not be practically feasible in real-world applications. On the other hand, work in the field of statistical analysis for enhancing data quality tends to focus on calculating dataset quality at a finer granularity within a specific dimension [38,48]. Quality is multidimensional, and solely relying on a single dimension for analysis can provide a limited perspective. These two articles are both based on the intrinsic attributes of the data and the contextual tasks for analysis [29,32]. However, the mathematical tools chosen in statistical analysis cannot achieve the goal of correctly assessing data quality. This paper divides data quality into multiple dimensions, allowing researchers to comprehensively assess the quality of data and identify areas in need of improvement. Since it focuses on enhancing data to improve model performance, it is essential to clearly define their definitions and relationships, while having effective methods for utilizing these dimensions to evaluate the effectiveness of data augmentation. So, we will compare our method with two baseline approaches: diversity and task independence calculated using the mean and min-max criteria. Since the criteria for quality fusion differ among these methods, we will employ our proposed quality fusion method to calculate the final scores for all the approaches. Baseline's formula is shown in Table 3.

Table 3. The formula of baseline.

Baseline	Metrics	Formula
mean criteria	$mean\_Q_1$	$1 - \frac{1}{N'^2} \sum_i^{N'} \sum_j^{N'} similarity(i,j)$
	$mean\_Q_6$	$\frac{1}{N' * N_t} \sum_i^{N'} \sum_j^{N_t} similarity(i,j)$
	$mean\_Q$	$mean\_Q_1 * mean\_Q_6$
minmax criteria	$minmax\_Q_1$	$1 - \frac{1}{N'} \sum_i^{N'} \min_{j \in D'} similarity(i,j)$
	$minmax\_Q_6$	$\frac{1}{N'} \sum_i^{N'} \max_{j \in D_t} similarity(i,j)$
	$minmax\_Q$	$minmax\_Q_1 * minmax\_Q_6$

In this section, this study conducted two sets of experiments to validate the effectiveness of the proposed method. The first set of experiments evaluated the correlation between the model precision on the test set and the quality evaluation results of the generated aug-

mented dataset using different parameters for data augmentation. Due to the enormous search space of data augmentation strategies, it was challenging to define it precisely. Therefore, this study randomly selected 7 sets of parameters as the comparative parameters for the experiment. The second set of experiments involved generating augmented datasets of different sizes using the same set of parameters for data augmentation. The performance of the model on these datasets was observed to see if it aligns with the algorithm results.

3.3. Evaluation Index

For the purpose of achieving higher model precision, this study adopts model evaluation metrics as the evaluation criteria for algorithm performance. After training on the augmented dataset, the model's precision is tested on the original test set. The model training parameter settings are shown in Table 4.

Table 4. The hyperparameters of DL model.

Dataset	Model	Hyperparameters
CIFAR10	densenet161	Epochs = 90, Init lr = 0.1 divide by 5 at 40th, 60th, 80th, Batchsize = 256, Weight decay = 5×10^{-4}, momentum = 0.9
CUB200	NtsNet	Epochs = 50, Lr = 0.001, Batchsize = 16, Weight decay = 1×10^{-4}, Momentum = 0.9

Due to the different dimensions between the scores of the method and the baseline and the accuracy of the model, we first divide all the model accuracies by the accuracy of the model on the original training set, similar to how the scores are calculated. We refer to this as the actual score of the augmented training set. Then, we consider using Cosine Similarity (CS) to calculate the similarity between the estimated scores and the actual scores. We sort the scores into a vector according to certain rules, such as enhancement magnitude or scale, and then calculate the cosine value of the angle between them. This cosine value reflects the similarity of the changing trends between the estimated scores and the actual scores. However, we still need to know the absolute distance between the two scores. Therefore, we also select Mean Squared Error (MSE) as our second metric. By combining these two metrics, we can comprehensively analyze the strengths and weaknesses of the algorithm. A larger CS and a smaller MSE indicate better results.

3.4. Ablation Study

This paper introduces an evaluation method for assessing the quality of data augmentation in an image dataset, aiming to better select augmentation techniques. In order to demonstrate the effectiveness of our proposed method in evaluating data augmentation based on local independence and global correlation information, as well as to validate our pipeline design, we conducted two ablation experiments. In Experiment 1, we focused on metric adjustment to confirm the observed differences between method results and estimation results from a statistical perspective. The Spearman's rank correlation coefficient contains a p-value to show the level of significance. The technique was statistically different when and only when the P value was below 0.05. In Experiment 2, we replaced the components in the framework to evaluate their impact on the results. Both ablation experiments were conducted on the CIFAR-10 dataset, and the experimental setup described earlier was used to train the network from scratch.

3.5. Case Study

In general, because different data types and different tasks require the machine learning model to learn different content, there is no absolute universal method for evaluating data augmentation quality. This article sets up experimental datasets from the perspectives of data types and downstream tasks. We will apply the methods described in this paper to two different datasets selected from various domains and tasks: EuroSAT and IMDB.

We have chosen datasets from significantly different domains—images and text, which represent two major data types. We also ensure that the datasets are large enough to facilitate meaningful augmentation and analysis. By calculating the score differences after augmentation using ADQE, we evaluate and visualize the differences between the two datasets with the largest score differences. EuroSAT, being an image dataset, undergoes the same augmentation and evaluation methods as described in this paper. For the text dataset IMDB, we apply data augmentation methods such as Optical Character Recognition (OCR), semantic augmentation, and summarization. Since the data type is different, we cannot directly use ADQE to calculate evaluation scores, and adjustments need to be made to the methods. The text dataset also involves extracting feature vectors to calculate diversity and task relevance, but text is composed of words rather than pixels. Q_2 and Q_3 need to be recalculated using words and characters. We combine these two metrics and redefine them as the information entropy calculated from frequency of the top 10,000 most frequent words.

4. Results and Discussion

4.1. Results of Comparative Experiments

The experimental results of the first group are presented in Figure 2. Figure 2 primarily illustrates the performance of our method compared to state-of-the-art methods on two different datasets, including the visual comparison and statistical analysis between the evaluation scores and model accuracy, as well as the variation trend of scores under different data augmentations. Figure 2a,b show the experimental results based on the CIFAR-10 dataset, while Figure 2c,d display the experimental results based on the CUB-200 dataset. From the experimental results, it can be observed that the entropy-based method provides the most similar quality results for all augmented datasets, which is further validated by the model's accuracy.

In Figure 2a,b, the entropy-based method shows the best fit to the model accuracy curve. Meanwhile, in Figure 2e,f, our results are validated using statistical analysis. We achieved excellent scores in both indicators, with CS approximately equal to 1 and MSE close to 0, indicating that entropy effectively captures the data characteristics and estimates corresponding evaluation scores. On the other hand, the evaluation results of methods like the minmax criteria continuously increase with the magnitude of data augmentation, deviating significantly from the trend of model accuracy. This is likely due to the noise introduced by data augmentation, leading to an overestimation of the quality of augmented datasets using the minmax criteria. This principle overlooks the overall distribution of the dataset and approximates its quality by statistical extreme values, resulting in a significant decrease in evaluation accuracy. Among all the figures, the minmax criteria consistently performs the worst. This observation is also reflected in the evaluation results based on the mean criteria in Figure 2b. The evaluation under overly strong data augmentation also overestimates the quality of the augmented dataset. In CUB-200, where the original data are scarce and very similar, with backgrounds mostly consisting of single skies, oceans, and forests, and birds differing slightly in feather color and shape, data augmentation significantly increases the diversity of the dataset. Changing a large portion of data distribution without perfectly matching the actual data distribution amplifies the diversity evaluation scores. The CS drops by one percentage point, and the MSE increases from below 0.1 to 1.1.

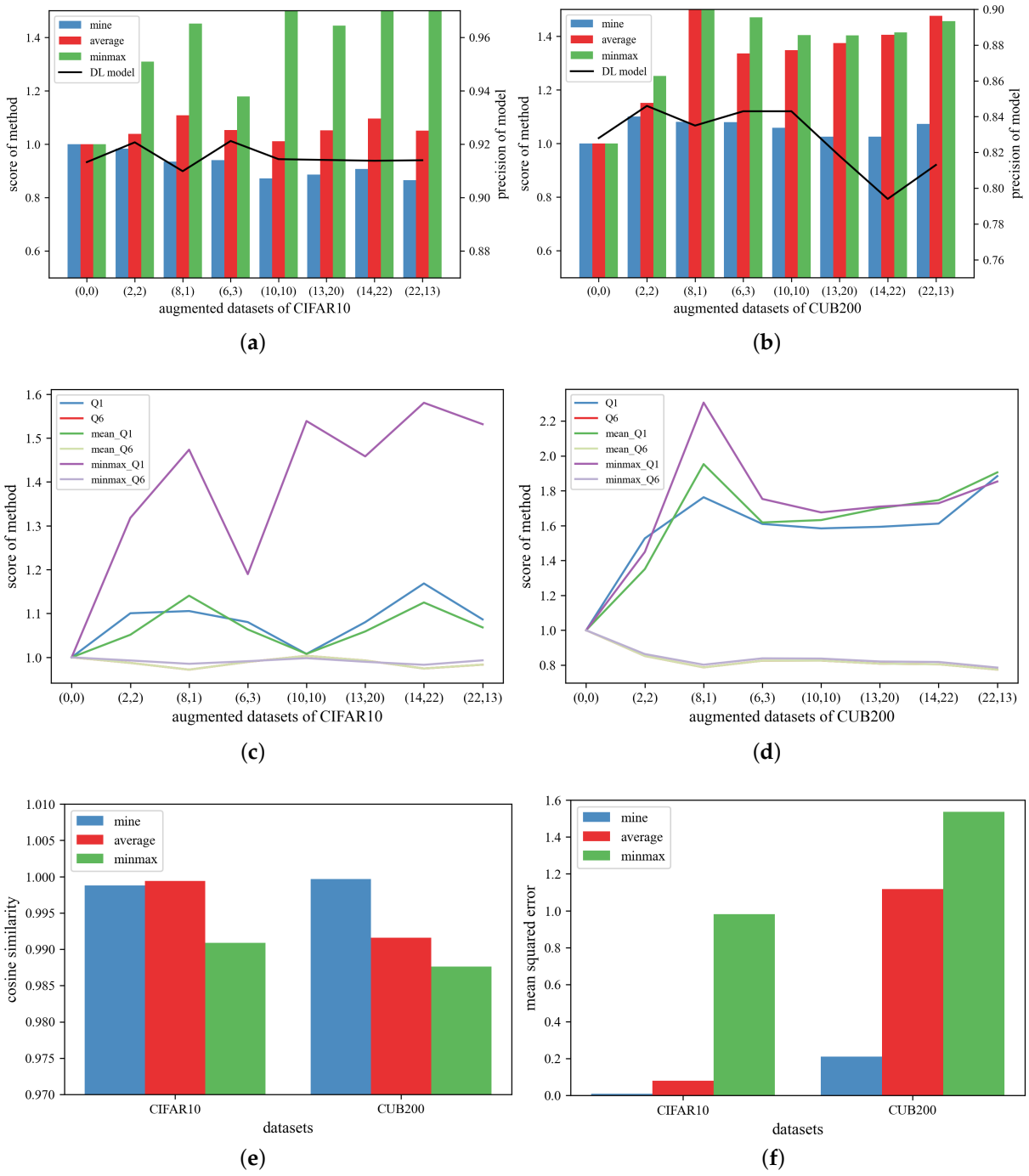


Figure 2. Under the same scale and different data augmentation scenarios, our method achieves the best quality evaluation results. (**a**) Evaluation of three algorithms and the model on the augmented dataset using the CIFAR-10 dataset. (**b**) Evaluation of three algorithms and the model on the augmented dataset using the CUB-200 dataset. (**c**) Partial indicator scores of the augmented dataset by three algorithms on the CIFAR-10 dataset. (**d**) Performance evaluation of the three algorithms using CS. (**e**) Performance evaluation of the three algorithms using MSE. (**f**) Partial indicator scores of the augmented dataset by three algorithms on the CUB-200 dataset. In (**a**–**c**,**f**), the x-axis represents the data augmentation strategies, including the selection probability of data augmentation and the magnitude of data augmentation. The selection of these strategies is randomly chosen within the given range. The word "mine" represents the methodology of this paper.

In our data augmentation quality evaluation metric calculation, we partition the overall evaluation metric based on the mean and variance of the data augmentation quality. The mean measures the distance between the augmented data and the target data distribution. The variance measures the distribution uniformity and diversity of the augmented data. From the q_6 curves in Figure 2c,d, for all methods, their mean calculation results are relatively consistent. They accurately calculate the distance between all augmented training data and the test set distributions. However, only the entropy method meets the criteria for variance estimation. Since semantic information has multiple dimensions and is not entirely related to labels, it is divided into class balance and diversity. Even through augmentation, class-related semantics will not be changed or lost. By using clustering dimensionality reduction, we can clearly understand the spatial distribution of the data, which remains relatively unchanged before and after augmentation. However, within the class, due to color changes, deformations, inversions, and other operations, these noises are substantially supplemented. We need to assess the distribution balance of data label-related features, and so on. The minmax criteria emphasizes the farthest distance of the data distribution. Although the mean criteria is more balanced, it is also affected by extreme values. Moreover, data augmentation methods are likely to inject many extreme values due to their randomness, thus affecting the evaluation. Entropy can smooth these extreme values, categorizing them together, and calculate the overall diversity evaluation value by statistically counting the effective number of categories under different dimensions. Experimental results demonstrate that our framework can effectively evaluate data augmentation quality by incorporating entropy, visually and comprehensively showcasing the improvements brought by data augmentation to the dataset.

From another perspective, Figure 2a–d illustrate that the model accuracy is significantly enhanced primarily through the increased diversity of the data. However, as the diversity increases, there is a bottleneck, and the task relevance decreases, resulting in poorer performance than the initial state. This is primarily related to the data augmentation techniques chosen in this paper. Most of the changes involve alterations in color space, shape, and orientation, which enhance data diversity and model robustness. Fewer enhancements focus on image quality or denoising, reducing the degree of association between samples in the dataset and the task objectives, which increases the dimensions of the data that need to be analyzed. The target variables become less understandable and predictable. Therefore, the experimental results inevitably show a decrease in task relevance and an increase in diversity. The intensity of data enhancement up to a certain point reduces the benefits and loses more data quality. By clearly defining dimensions, researchers can better guide the selection and implementation of data augmentation strategies. Using data quality dimensions to evaluate the effectiveness of data augmentation in experiments helps provide empirical evidence. This allows researchers to quantify improvements and demonstrate that the measures taken have indeed enhanced model performance.

The results of the second group of experiments are shown in Figure 3. The evaluation trends of the three methods are mostly consistent, showing an improvement in the quality fusion score as the scale increases, although the rate of improvement decreases. The minmax principle still exaggerates the quality of data augmentation. The model precision also increases as the scale expands, but gradually starts to decline after reaching a four-fold scale. This indicates that the repetition of data augmentation-generated images can have a negative impact on model performance, leading to overfitting. It demonstrates that more data are not always better. According to Figure 3b, our algorithm and the averaging method have similar performance, which is superior to the minmax values.

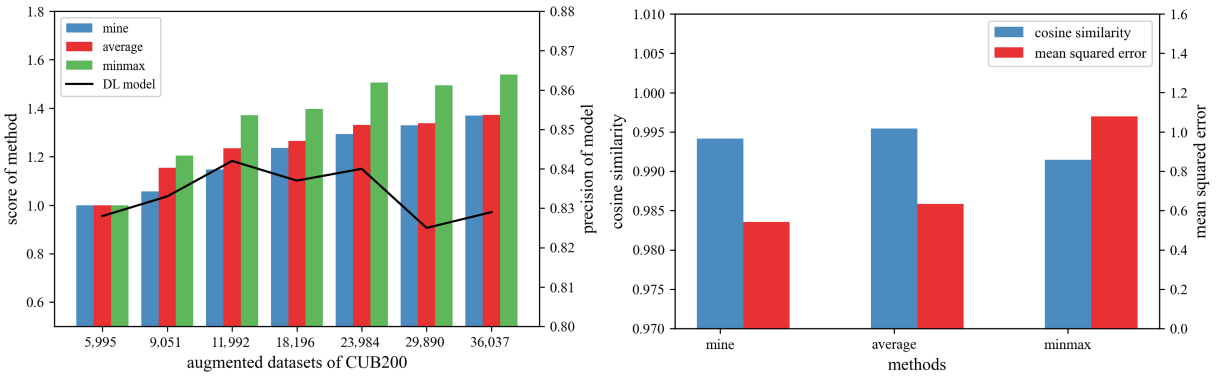


Figure 3. Under the same data augmentation with different scales, our algorithm and the averaging method exhibit similar performance. (**a**) Evaluation of the three algorithms and the model on the augmented dataset of CUB-200 dataset. (**b**) Performance evaluation of the three algorithms using CS and MSE. The word "mine" represents the methodology of this paper.

4.2. Results of Ablation Experiments

Table 5 presents four methods for metrics fusion. In the case of calculating metrics using multiplication, if we continue to use multiplication for index weights, it would lose the balancing effect of importance. Therefore, we chose to use power operations to further enhance the balancing role of parameters. For each fusion method, we evaluate the accuracy of the evaluation results using CS and MSE. The fusion method used in this study achieved the best scores in all four indicators. The results in the second and third rows are one order of magnitude higher than those in the first and fourth rows, with MSE increasing from 0.1357 and 0.5363 to 5.6027 and 9.7748. This also confirms that in the estimation of variance, the labels and semantic cannot be simply regarded as two sets of linearly related variables. The semantic information needs further decomposition, and the final scores need to be fused using addition for the unrelated parts. The first and fourth rows achieved approximately the same excellent scores, with CS differing by less than 1 percentage point. However, in the CUB dataset, the MSE differed by three times. The reason is the neglect of the correlation between variance and mean. In large and relatively balanced datasets like CIFAR-10, this effect is not evident. But for small datasets, data augmentation can easily disrupt data semantics and introduce irrelevant noise, affecting model training. This is reflected in the increase in variance evaluation results while the mean evaluation results decrease. Therefore, using addition to fuse variance and mean cannot well reflect the relationship between the two indicators.

In Tables 6 and 7, we can visually observe the correlations between indicators. One notable observation is that the correlation between Q_1 and Q_6 is approximately -1. This confirms that there is indeed a strong negative correlation between them, consistent with the previous discussions. Surprisingly, we also found high correlations between Q_1 and Q_4, as well as between Q_2, Q_3, and Q_5. The high correlation between Q_1 and Q_4 may be due to the random generation of augmented datasets. It results in different numbers of samples for each class, causing Q_4 and Q_1 to increase synchronously and be considered highly correlated. If the size of tests is larger and more diverse, this correlation will gradually decrease. The high correlation between Q_2, Q_3, and Q_5 may be due to manual rule intervention, which leads to a more uniform color space distribution in the dataset and reduces the use of color as a criterion for unsupervised clustering. This improves the accuracy of clustering results. This also indirectly demonstrates that data augmentation indeed helps the model learn from data.

Table 5. Results of ablation experiments.

Fusion of Quality Metrics	Formula	CS_{cub}	MSE_{cub}	CS_{cifar}	MSE_{cifar}
$(Q_{Diversity} + Q_{ClassBalance}) \times Q_{TaskRelevance}$	$\frac{Q_6'}{Q_6} \times \sum_{i=1}^{5} w_i \frac{Q_i'}{Q_i} / \sum_{i=1}^{5} w_i$	**0.9999** *	**0.1357**	**0.9997**	**1.4841**
$Q_{Diversity} \times Q_{ClassBalance} \times Q_{TaskRelevance}$	$\prod_{i=1}^{6} (\frac{Q_i'}{Q_i})^{r_i}$	0.9805	5.6027	0.9721	16.5365
$Q_{Diversity} \times (Q_{ClassBalance} + Q_{TaskRelevance})$	$\prod_{i=1}^{3} (\frac{Q_i'}{Q_i})^{r_i} \times \sum_{i=4}^{6} w_i \frac{Q_i'}{Q_i} / \sum_{i=4}^{6} w_i$	0.9727	9.7748	0.9916	98.9535
$Q_{Diversity} + Q_{ClassBalance}) + Q_{TaskRelevance}$	$\sum_{i=1}^{6} w_i \frac{Q_i'}{Q_i} / \sum_{i=1}^{6} w_i$	0.9977	0.5363	0.9995	1.8736

* The bold values are the best.

Table 6. The correlation coefficient of each metric in CIFAR-10.

Metrics	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
Q_1	1					
Q_2	−0.05	1				
Q_3	−0.29	0.76 *	1			
Q_4	0.81 *	0.10	−0.10	1		
Q_5	−0.19	0.62 *	0.81 *	−0.14	1	
Q_6	−0.92 **	0.02	0.24	−0.57	0.07	1

* $p < 0.05$, ** $p < 0.01$.

Table 7. The correlation coefficient of each metric in CUB-200.

Metrics	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
Q_1	1					
Q_2	−0.23	1				
Q_3	−0.71 *	0.73 *	1			
Q_4	0.76 *	0.11	−0.24	1		
Q_5	−0.71 *	0.79 *	0.95 **	−0.19	1	
Q_6	−0.97 **	0.36	0.79 *	−0.66	0.8 *	1

* $p < 0.05$, ** $p < 0.01$.

However, there are slight differences in the experimental results between the two datasets. For example, in CIFAR-10, Q_6 is strongly correlated only with Q_1. But in CUB-200, Q_6 is strongly correlated with not only Q_1 but also Q_3 and Q_5. In particular, the correlation with Q_5 increased from 0.07 to 0.8, showing two extremes. From the perspective of dataset properties, CIFAR-10 is a large-scale dataset with data from various scenarios and weather conditions, and data augmentation has a limited impact on its basic image feature distribution, and the distribution of each class changes little. Data augmentation mainly affects semantic information in this dataset, such as object inversion not affecting recognition. Therefore, only Q_1 and Q_6 show correlation. However, in the small dataset CUB-200, most images have a single color and similar backgrounds. Image features have a greater impact on the final results, which is also reflected in the correlation of indicator results. Therefore, when using our framework to evaluate data augmentation quality, analyzing the correlations of various indicators can help understand the strengths and weaknesses of the dataset.

Figure 4 illustrates the impact of clustering algorithms on the framework results, where only Q_1 and Q_6 are accelerated by clustering in the framework. Figure 4a,b show the accuracy of the results before and after clustering. The original algorithm uses brute force to compute the distance between pairwise image vectors to obtain evaluation results. In Figure 4a, the clustered results show some differences compared to the original results, with a similar trend but a decrease in evaluation scores after clustering. However, as Figure 4b indicates, both indicators improve after clustering, proving that clustering actually helps improve the accuracy of the evaluation results. Figure 4c displays the saved running time through clustering. The x-axis represents the product of the number of classes and the

number of samples in each class in the dataset. It is evident that when the number of classes is large and the number of samples per class is small, clustering only saves about half of the time. However, when the number of samples per class is much larger than the number of classes, clustering can save over 90% of the time. This is because Q_1 separates the dataset based on classes, and the time complexity within each class is $O(n^2)$. After fast clustering, since the number of classes is roughly the same, it can be considered a constant factor. Therefore, the fewer the number of classes and the more samples per class, the more time can be saved. In Figure 5, it can be observed that replacing different pre-trained models does not significantly affect the evaluation results of this method. The more sufficient the pre-training, the more accurate the grasp of image features, and the generated image feature vectors also have certain discriminative power. The two pre-trained models selected in this study, both trained on ImageNet, did not show significant differences.

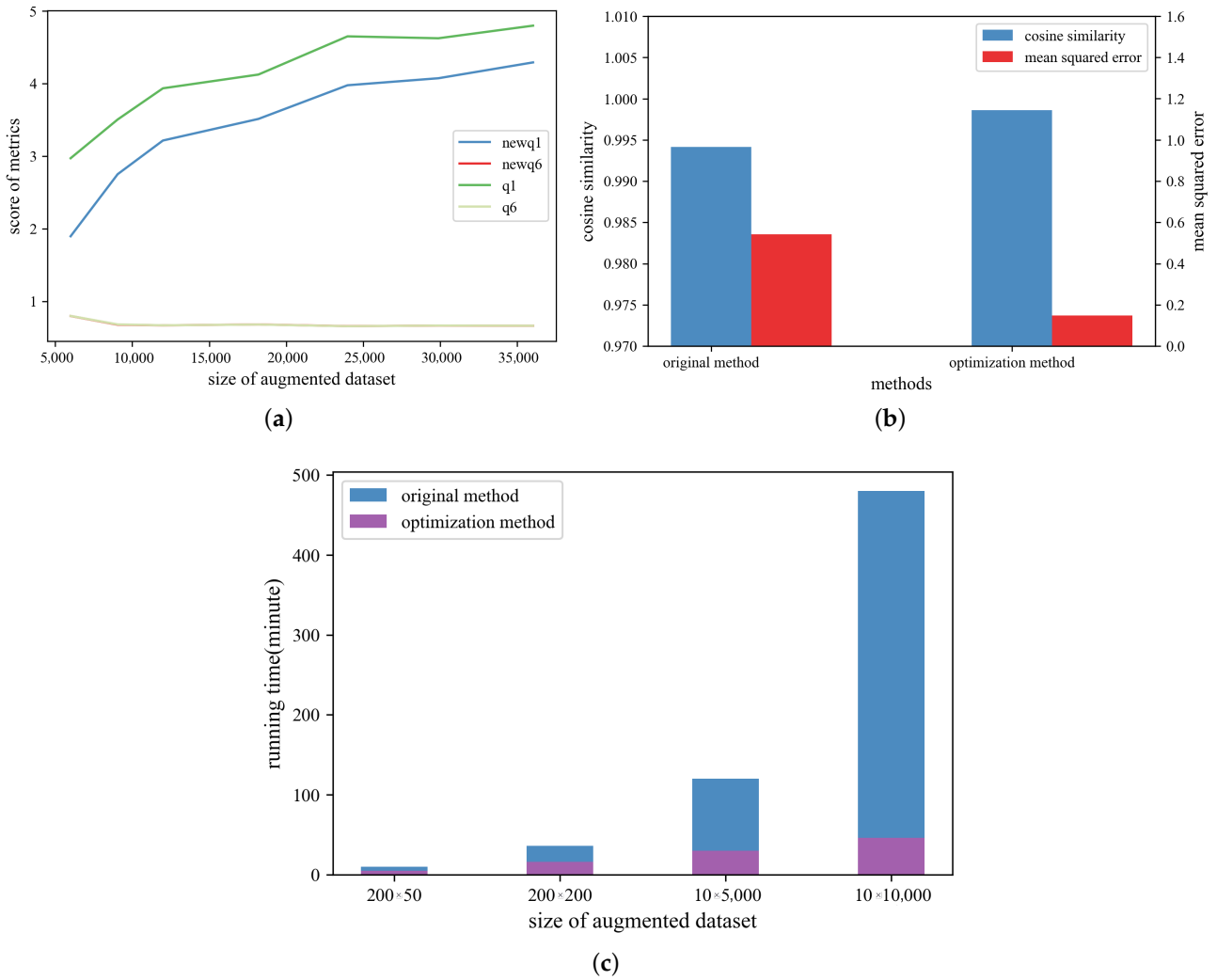
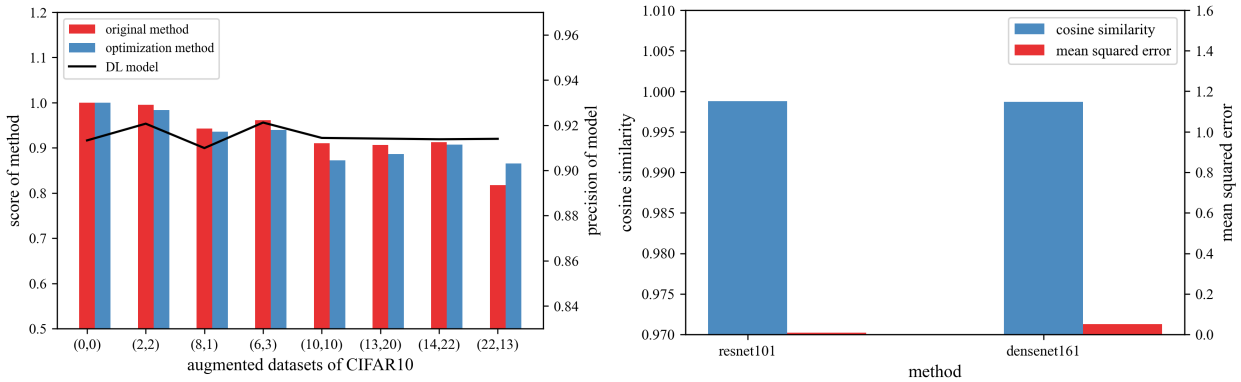


Figure 4. The clustering results have not only optimized the computation of the algorithm but also improved its performance. (**a**) The running times of the unoptimized and optimized algorithms are shown for different numbers of categories and samples within each category. The x-axis represents the product of the number of categories and the number of samples within each category in the dataset. (**b**) Partial scores of the original algorithm and the optimized algorithm are presented for the CUB-200 dataset. The x-axis represents the total number of samples in the augmented training set. (**c**) Performance evaluation of the original algorithm and the optimized algorithm in the CUB-200 dataset.

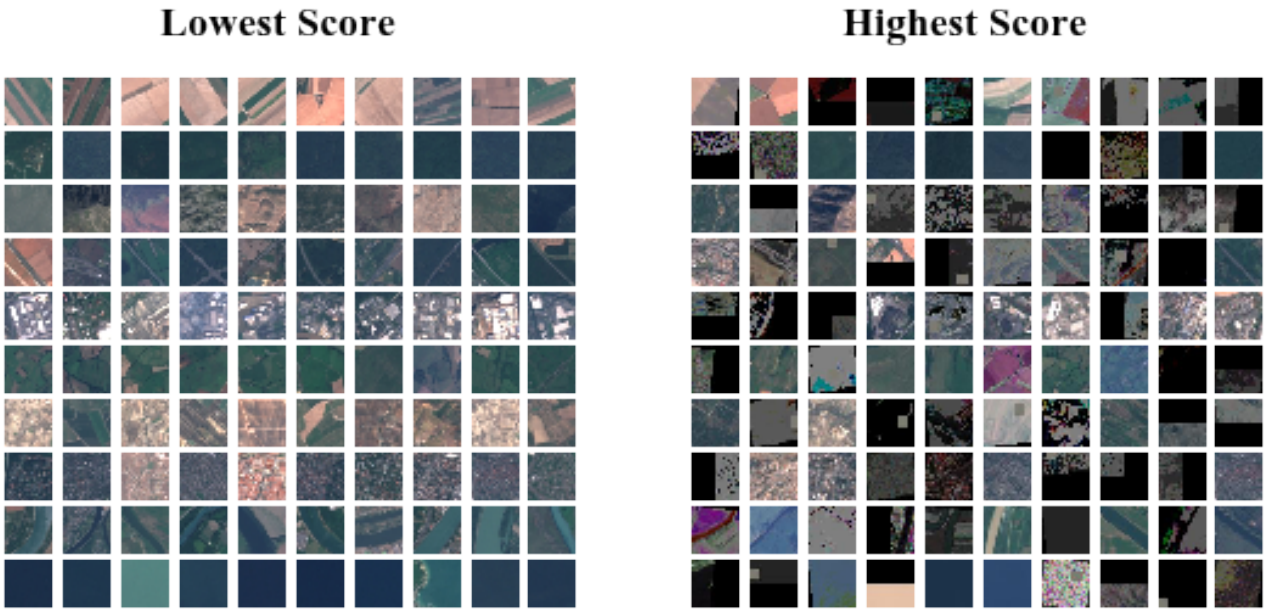


(**a**) (**b**)

Figure 5. Under the same data augmentation with different scales, our algorithm and the averaging method exhibit similar performance. (**a**) Evaluation of the three algorithms and the model on the augmented dataset of CUB-200 dataset. (**b**) Performance evaluation of the three algorithms using CS and MSE.

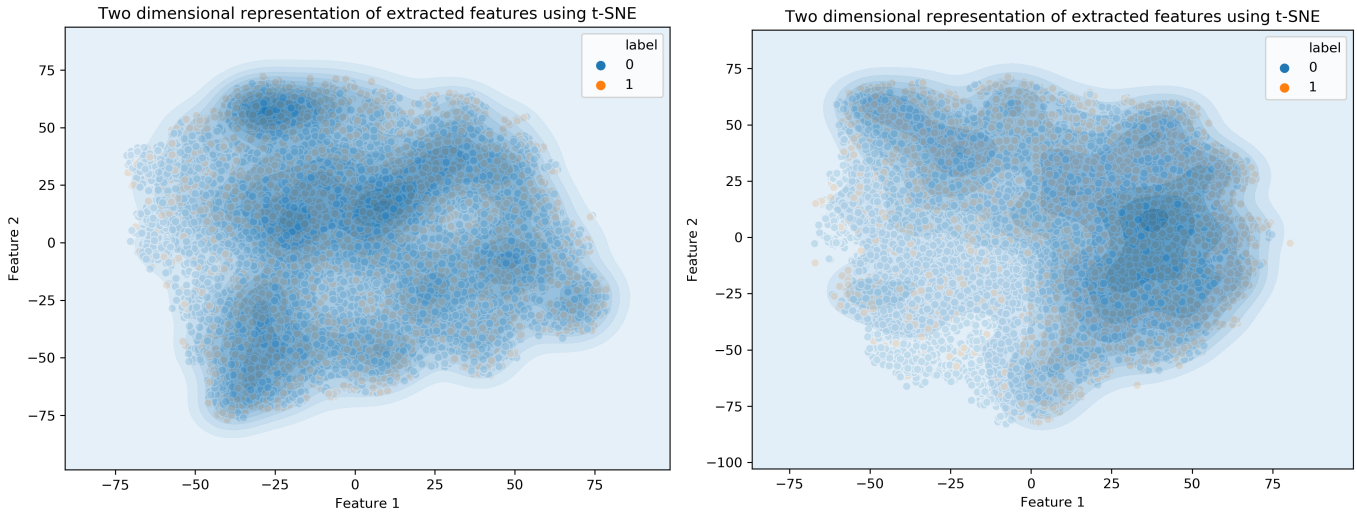
4.3. Applicability and Potential Applications

In Figures 6 and 7, we assessed the quality of each dataset and visualized the scores for the original dataset and the highest dataset. It can be seen that from a visual perspective, there are significant differences between the two. Figure 6 presents a comparison of the EuroSAT dataset. On the left, the original images within the same category mostly share a similar color and content. On the right, due to the effects of data augmentation, there is increased diversity in color and brightness. However, in some images, it becomes difficult to discern their content with the naked eye. Figure 7 illustrates the dimensionality reduction visualization of the IMDB dataset using the t-SNE algorithm. It is evident that the dataset on the right, which has higher scores, exhibits better data separation. Most of the positive data points are concentrated on the right. This indicates that the classes have good separability, which can effectively enhance the model's performance.



(**a**) (**b**)

Figure 6. This is a display of the EuroSAT dataset, each row is a different class. (**a**) Origin dataset. (**b**) The augmented dataset with the highest score.



(**a**) (**b**)

Figure 7. The t-SNE algorithm reduces dimensionality of all features in the IMDB dataset and visualizes them. (**a**) Origin dataset. (**b**) The augmented dataset with the highest score.

The context of intrinsic data attributes and the intended use of the data in the evaluation method cannot be balanced well, which limits its applicability. For example, remote sensing images inherently have issues like low contrast, noise, and blurriness. Additionally, remote sensing images often contain imprecisely shaped objects, unlike medical imaging, which require precise target recognition. Therefore, remote sensing images are not highly sensitive to task relevance, and optimizing intrinsic data attributes for quality improvement results in more significant quality loss compared to the decrease in task relevance. In contrast, medical imaging has sufficiently high image quality. In an attempt to enhance diversity, it leads to a loss in data quality and task relevance, ultimately resulting in an overall decline in quality. Different tasks have varying quality requirements for datasets, and during the calculation, parameter variations in three areas—diversity, class balance, and task relevance—need to be considered. This is also a direction we need to focus on in the future. When our method is not tuned to the correct parameter resulting in poor performance, a more well-performing parameter can be fitted by combining the evaluation results of multiple augmented datasets with a small number of training results. Then try evaluating a new augmented dataset again until the parameter is similar to most of the training results.

Different downstream tasks have different priorities. The evaluation method may succeed for the assumed task but fail for the target task. For example, in the case of object recognition tasks, the optimal augmented dataset obtained using the method described in this paper may be suboptimal for this task. This is because image similarity metrics are better suited for classification tasks rather than object recognition, where the goal is to detect similarity between small regions. Task relevance metrics, on the other hand, can replace this by calculating the similarity between annotated regions and other regions.

Data can exhibit significant differences in terms of structure, format, complexity, size, noise levels, and more. Evaluation methods tailored to one data type may fail when applied to another data type. For image datasets, we calculate statistics on pixels and textures. However, text datasets are composed of characters, words, and sentences, so the intrinsic data quality of different data types needs rules established by domain experts for measurement. In addition to this, this paper's method reflects high applicability and robustness in data diversity. The main reason is that no matter what type of deep learning data can use neural networks to extract multi-dimensional feature vectors, which constitutes most of the indicators of this paper's method based on the calculation of feature vectors, so to a certain extent this paper has the ability to adapt to most of the fields.

In most real-world datasets, a certain degree of anomalies, noise, errors, and outliers can be expected. Truly clean and pristine data are a rare find. This is particularly true for sensor data, remote sensing data, and internet-derived data, which often exhibit higher variability and a higher incidence of anomalies. The process of data augmentation can introduce new error patterns. In this paper, we split the dataset into three attributes and utilize techniques such as information entropy to effectively quantify the number of patterns and task relevance. Excessive errors can lead to a decrease in task relevance, which can manifest in the final results. However, because this paper employs unsupervised algorithms, it is challenging to distinguish and optimize true anomalies from acceptable variations.

The diversity and complexity of real-world data make the development of universally applicable data quality assessment methods inherently challenging. Thoughtful method design, extensive evaluation, avoiding overfitting, and relaxing assumptions can enhance applicability. In practical scenarios, the methods outlined in this paper can guide the construction of datasets. If the collected sample dataset is insufficient for training deep learning models, you can evaluate the effectiveness of data augmentation using quantitative metrics defined based on the three dataset attributes and the ultimate goal, such as improving classification accuracy, reducing error rates, or enhancing signal-to-noise ratios. Utilizing data augmentation and assessment methods can address the long-tail problem in data, improve data consistency, and reduce the labeling workload. As more real-world data become available, it is essential to continuously reassess and enhance data augmentation. The demand for augmentation may evolve over time. In summary, the methods described in this paper contribute to the development of well-generalized models from limited and imperfect real-world data.

Furthermore, the methods outlined in this paper are easy to implement and can be seamlessly integrated into the deep learning workflow using popular frameworks like TensorFlow or PyTorch. Leveraging the data pipelines within these frameworks and the data monitoring integrated into them, it becomes straightforward to quickly compute intrinsic data attribute metrics and feature vectors generated by pre-trained models after generating multiple augmented datasets from the input data. Once the optimal dataset for evaluation is obtained, you can proceed directly to training your model.

5. Conclusions

This study aims to enhance model performance through the improvement of data quality. By categorizing data quality into three key dimensions, diversity, class balance, and task relevance, and evaluating the effectiveness of data augmentation within these dimensions, we have achieved the following key outcomes and conclusions.

Firstly, we have successfully deconstructed the complexity of data quality into three essential dimensions. This aids in providing a more comprehensive understanding of data quality. Diversity ensures the inclusion of various sample types in the dataset, class balance helps address imbalances in class distribution, and task relevance ensures the alignment of data with the actual task at hand.

Secondly, by assessing the impact of data augmentation methods across these three dimensions, we can quantitatively measure the influence of different enhancement strategies on data quality. Our experimental results demonstrate that, with reasonable selection and adjustment of augmentation strategies, significant improvements can be made in data diversity and class balance while maintaining a high degree of relevance to the task.

Most importantly, our work holds significant practical implications. In the modern fields of machine learning and artificial intelligence, data serves as the foundation of successful models. By elevating data quality, we can enhance model generalization, mitigate overfitting risks, improve model robustness in real-world scenarios, and provide more accurate predictive and decision-making support across various application domains. This impact extends to critical areas such as medical diagnostics, financial risk analysis, autonomous driving, and beyond.

In the future, we aim to further explore the relationships among data quality dimensions and strive towards the automatic selection of parameters tailored to specific domains and tasks. The most notable improvement will be in terms of efficiency, as manual parameter selection and adjustment will no longer be necessary. Additionally, this approach will reduce subjectivity and bias, ensuring the replicability and comparability of experimental results. Most importantly, it will simplify experimentation, making this method accessible to a broader range of researchers interested in understanding the principles of data construction.

In summary, our research provides a systematic approach to enhancing data quality, enabling researchers to better comprehend, evaluate, and enhance data for improved machine learning model performance. This work offers robust guidance for future research and applications, with the potential to make a positive impact in data-driven fields.

Author Contributions: Conceptualization, X.C., Y.L. and C.M.; methodology, X.C. and Y.L.; software, X.C., Y.L. and C.M.; validation, X.C., Y.L. and C.M.; formal analysis, X.C. and Y.L.; investigation, Y.L.; resources, X.C., Y.L., Z.X. and C.M.; data curation, X.C., Y.L., H.L. and S.Y.; writing—original draft preparation, X.C. and Y.L.; visualization, X.C. and Y.L.; supervision, X.C., Z.X. and C.M.; project administration, X.C. and C.M.; funding acquisition, X.C. and C.M.; X.C., Y.L. and C.M.: significance contributions to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Outstanding Youth Team Project of Central Universities(QNTD202308) and National Key R&D Program of China (2022YFF1302700).

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, T.; Chen, J.; Li, F.; Zhang, K.; Lv, H.; He, S.; Xu, E. Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. *ISA Trans.* **2022**, *119*, 152–171. [PubMed]
2. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef] [PubMed]
3. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362*, 1140–1144. [CrossRef] [PubMed]
4. Hao, X.; Liu, L.; Yang, R.; Yin, L.; Zhang, L.; Li, X. A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition. *Remote Sens.* **2023**, *15*, 827. [CrossRef]
5. Chen, Y.; Yang, X.H.; Wei, Z.; Heidari, A.A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q. Generative adversarial networks in medical image augmentation: A review. *Comput. Biol. Med.* **2022**, *144*, 105382. [CrossRef]
6. Yang, J.; Guo, X.; Li, Y.; Marinello, F.; Ercisli, S.; Zhang, Z. A survey of few-shot learning in smart agriculture: Developments, applications, and challenges. *Plant Methods* **2022**, *18*, 28. [CrossRef]
7. Maslej-Krešňáková, V.; Sarnovskỳ, M.; Jacková, J. Use of Data Augmentation Techniques in Detection of Antisocial Behavior Using Deep Learning Methods. *Future Internet* **2022**, *14*, 260. [CrossRef]
8. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text data augmentation for deep learning. *J. Big Data* **2021**, *8*, 101. [CrossRef]
9. Gong, C.; Wang, D.; Li, M.; Chandra, V.; Liu, Q. Keepaugment: A simple information-preserving data augmentation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1055–1064.
10. Iwana, B.K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*, e0254841. [CrossRef]
11. Zhou, X.; Hu, Y.; Wu, J.; Liang, W.; Ma, J.; Jin, Q. Distribution bias aware collaborative generative adversarial network for imbalanced deep learning in industrial IoT. *IEEE Trans. Ind. Inform.* **2022**, *19*, 570–580. [CrossRef]
12. Bishop, C.M. Training with noise is equivalent to Tikhonov regularization. *Neural Comput.* **1995**, *7*, 108–116. [CrossRef]
13. Hernández-García, A.; König, P. Data augmentation instead of explicit regularization. *arXiv* **2018**, arXiv:1806.03852.
14. Carratino, L.; Cissé, M.; Jenatton, R.; Vert, J.P. On mixup regularization. *arXiv* **2020**, arXiv:2006.06049.
15. Shen, R.; Bubeck, S.; Gunasekar, S. Data augmentation as feature manipulation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 19773–19808.

16. Ilse, M.; Tomczak, J.M.; Forré, P. Selecting data augmentation for simulating interventions. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 4555–4562.
17. Allen-Zhu, Z.; Li, Y. Feature purification: How adversarial training performs robust deep learning. In Proceedings of the 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), Denver, CO, USA, 7–10 February 2022; pp. 977–988.
18. Kong, Q.; Chang, X. Rough set model based on variable universe. *CAAI Trans. Intell. Technol.* **2022**, *7*, 503–511. [CrossRef]
19. Zhao, H.; Ma, L. Several rough set models in quotient space. *CAAI Trans. Intell. Technol.* **2022**, *7*, 69–80. [CrossRef]
20. Kusunoki, Y.; Błaszczyński, J.; Inuiguchi, M.; Słowiński, R. Empirical risk minimization for dominance-based rough set approaches. *Inf. Sci.* **2021**, *567*, 395–417. [CrossRef]
21. Chen, S.; Dobriban, E.; Lee, J.H. A group-theoretic framework for data augmentation. *J. Mach. Learn. Res.* **2020**, *21*, 9885–9955.
22. Mei, S.; Misiakiewicz, T.; Montanari, A. Learning with invariances in random features and kernel models. In Proceedings of the Conference on Learning Theory, Boulder, CO, USA, 15–19 August 2021; pp. 3351–3418.
23. Wand, Y.; Wang, R.Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* **1996**, *39*, 86–95. [CrossRef]
24. Abdullah, M.Z.; Arshah, R.A. A review of data quality assessment: Data quality dimensions from user's perspective. *Adv. Sci. Lett.* **2018**, *24*, 7824–7829. [CrossRef]
25. Firmani, D.; Mecella, M.; Scannapieco, M.; Batini, C. On the meaningfulness of "big data quality". *Data Sci. Eng.* **2016**, *1*, 6–20. [CrossRef]
26. Jarwar, M.A.; Chong, I. Web objects based contextual data quality assessment model for semantic data application. *Appl. Sci.* **2020**, *10*, 2181. [CrossRef]
27. Sim, K.; Yang, J.; Lu, W.; Gao, X. MaD-DLS: Mean and deviation of deep and local similarity for image quality assessment. *IEEE Trans. Multimed.* **2020**, *23*, 4037–4048. [CrossRef]
28. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [CrossRef]
29. Chen, H.; Chen, J.; Ding, J. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Trans. Reliab.* **2021**, *70*, 831–847. [CrossRef]
30. Gosain, A.; Saha, A.; Singh, D. Measuring harmfulness of class imbalance by data complexity measures in oversampling methods. *Int. J. Intell. Eng. Inform.* **2019**, *7*, 203–230. [CrossRef]
31. Bellinger, C.; Sharma, S.; Japkowicz, N.; Zaïane, O.R. Framework for extreme imbalance classification: SWIM—Sampling with the majority class. *Knowl. Inf. Syst.* **2020**, *62*, 841–866. [CrossRef]
32. Li, A.; Zhang, L.; Qian, J.; Xiao, X.; Li, X.Y.; Xie, Y. TODQA: Efficient task-oriented data quality assessment. In Proceedings of the 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), Shenzhen, China, 11–13 December 2019; pp. 81–88.
33. Delgado-Bonal, A.; Marshak, A. Approximate entropy and sample entropy: A comprehensive tutorial. *Entropy* **2019**, *21*, 541. [CrossRef]
34. Li, Y.; Chao, X.; Ercisli, S. Disturbed-entropy: A simple data quality assessment approach. *ICT Express* **2022**, *8*, 309–312. [CrossRef]
35. Liu, L.; Miao, S.; Liu, B. On nonlinear complexity and Shannon's entropy of finite length random sequences. *Entropy* **2015**, *17*, 1936–1945. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Sarfraz, S.; Sharma, V.; Stiefelhagen, R. Efficient parameter-free clustering using first neighbor relations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8934–8943.
38. Friedman, D.; Dieng, A.B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *arXiv* **2022**, arXiv:2210.02410.
39. Mishra, S.P.; Sarkar, U.; Taraphder, S.; Datta, S.; Swain, D.P.; Saikhom, R.; Panda, S.; Laishram, M. Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). *Int. J. Livest. Res.* **2017**, *7*, 60–78.
40. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.
41. Lore, K.G.; Akintayo, A.; Sarkar, S. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.* **2017**, *61*, 650–662. [CrossRef]
42. Yang, Y.; Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19290–19301.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. Xu, Y.; Lu, Y. Adaptive weighted fusion: A novel fusion approach for image classification. *Neurocomputing* **2015**, *168*, 566–574. [CrossRef]
45. Ahmad, S.; Pal, R.; Ganivada, A. Rank level fusion of multimodal biometrics using genetic algorithm. *Multimed. Tools Appl.* **2022**, *81*, 40931–40958. [CrossRef]
46. Nawaz, S.; Calefati, A.; Caraffini, M.; Landro, N.; Gallo, I. Are these birds similar: Learning branched networks for fine-grained representations. In Proceedings of the 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), Dunedin, New Zealand, 2–4 December 2019; pp. 1–5.

47. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
48. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.